An Alignment-Free Regression Approach for Estimating Allele-Specific Expression using RNA-Seq Data

Chen-Ping Fu, Vladimir Jojic, and Leonard McMillan

Department of Computer Science, University of North Carolina, Chapel Hill, NC
{ping, vjojic, mcmillan}@cs.unc.edu

Abstract. RNA-seq technology enables large-scale studies of allele-specific expression (ASE), or the expression difference between maternal and paternal alleles. Here, we study ASE in animals for which parental RNA-seq data are available. While most methods for determining ASE rely on read alignment, read alignment either leads to reference bias or requires knowledge of genomic variants in each parental strain. When RNA-seq data are available for both parental strains of a hybrid animal, it is possible to infer ASE with minimal reference bias and without knowledge of parental genomic variants. Our approach first uses parental RNA-seq reads to discover maternal and paternal versions of transcript sequences. Using these alternative transcript sequences as features, we estimate abundance levels of transcripts in the hybrid animal using a modified lasso linear regression model.

We tested our methods on synthetic data from the mouse transcriptome and compared our results with those of Trinity, a state-of-the-art *de novo* RNA-seq assembler. Our methods achieved high sensitivity and specificity in both identifying expressed transcripts and transcripts exhibiting ASE. We also ran our methods on real RNA-seq mouse data from two F1 samples with wild-derived parental strains and were able to validate known genes exhibiting ASE, as well as confirm the expected maternal contribution ratios in all genes and genes on the X chromosome.

Keywords: Allele-Specific Expression, RNA-seq, Lasso Regression

1 Introduction

Recent advances in high-throughput RNA-seq technology have enabled the generation of massive amounts of data for investigation of the transcriptome. While this offers exciting potential for studying known gene transcripts and discovering new ones, it also necessitates new bioinformatic tools that can efficiently and accurately analyze such data.

Current RNA-seq techniques generate short reads from RNA sequences at high coverage, and the main challenge in RNA-seq analysis lies in reconstructing transcripts and estimating their relative abundances from millions of short (35-250 bp) read sequences. A common approach is to first map short reads onto

a reference genome, and then estimate the abundance in each annotated gene region. Such reference-alignment methods include TopHat [25], Cufflinks [27] and Scripture [10], which use algorithms such as the Burrows-Wheeler transform [1] to achieve fast read alignment. These methods are well established in the RNA-seq community and there exist many auxiliary tools [25] [26] for downstream analysis.

However, aligning reads to a reference genome has some disadvantages. First, read alignment assumes samples are genetically similar to the reference genome, and as a result, samples that deviate significantly from the reference frequently have a large portion of unmapped reads. This bias favors mapping reads from samples similar to the reference genome and is known as "reference bias." Second, alignment methods typically cannot resolve the origin of reads that map to multiple locations in the genome, resulting in reads being arbitrarily mapped or discarded from analysis. Suggested workarounds to the first problem of reference bias involve creating new genome sequences, typically by incorporating known variants, to use in place of the reference genome for read alignment [23]. However, this requires prior knowledge of genomic variants in the targeted RNA-seq sample, which is sometimes difficult and expensive to obtain.

Another class of methods perform *de novo* assembly of transcriptomes using *De Bruijn* graphs of k-mers from reads [7] [21]. These methods enable reconstruction of the transcriptome in species for which no reference genomic sequence is available. While these methods offer the possibility of novel transcript discovery, their *de novo* nature makes it difficult to map assembled subsequences back to known annotated transcripts. Furthermore, estimation of transcript expression levels in these methods is not straightforward and generally involves alignment of assembled contigs to a reference genome [7] [21], which reintroduces the possibility for reference bias.

Expression level estimation is particularly difficult for outbred diploid organisms, since each expressed transcript may contain two different sets of alleles, one from each parental haplotype. In some transcripts, one parental allele is preferentially expressed over another, resulting in what is known as allele-specific expression (ASE). It is often biologically interesting to identify genes and transcripts exhibiting ASE, as well as estimate the relative expression levels of the maternal and paternal alleles [8] [29]. Prior to the introduction of RNA-seq, ASE studies often relied on microarray technology. Although microarrays are able to identify genes exhibiting ASE, they generally examine a small number of genes, with expression level estimates in highly relative terms [19] [22]. The abundance of data from RNA-seq not only enables large-scale ASE studies incorporating the entire transcriptome, but also provides several means for direct estimation of more accurate expression levels, such as using alignment pile-up heights.

Current RNA-seq-based methods for analyzing ASE rely on reference transcriptome alignment [23] [24], which is again subject to reference bias and requires prior knowledge of genomic variants in the strains of interest. Reference bias is particularly problematic in ASE analysis, since it can falsely enhance relative expression in one parental strain over another.

In the case where RNA-seq data of all three members of a mother-father-child trio are available, we can utilize the RNA-seq data from the parental strains and eliminate the need for prior knowledge of their genomic variants. Here, we examine ASE in F1 mouse strains, which are first-generation offspring of two distinct isogenic parental strains. We separately construct maternal and paternal versions of transcripts using RNA-seq reads from the parental strains and annotated reference transcripts, creating a set of candidate transcript sequences the F1 strain could express. We then estimate the expression level of each candidate transcript in the F1 strain using a modified lasso regression model [11]. Lasso regression has been proposed by Li et al. [17] in the context of RNA-seq isoform expression level estimation, but not in the context of estimating ASE without reference alignment. We choose to use lasso regularization since it drives parameters to zero, enabling us to effectively eliminate non-expressed isoforms that have significant k-mer overlaps with expressed isoforms. We modify the lasso penalty slightly to prefer assigning higher F1 expression levels in transcripts with k-mers that appear frequently in the parental RNA-seq reads, due to the assumption that most highly expressed genes in the parents should also be highly expressed in the F1 strain.

We tested our methods on synthetic RNA-seq data from the wild-derived mouse strains CAST/EiJ and PWK/PhJ, along with F1 offspring CASTxPWK, with CAST/EiJ as the maternal strain and PWK/PhJ as the paternal strain. We also tested on real RNA-seq data from a CAST/EiJ, PWK/PhJ, CASTxPWK trio and a CAST/EiJ, WSB/EiJ, CASTxWSB trio, both using CAST/EiJ as the maternal strain. The CAST/EiJ, PWK/PhJ, and WSB/EiJ mouse strains are isogenic, and all three have well-annotated genomes that differ significantly from each other and from the mouse reference sequence [30], which is largely based on the C57BL/6J strain (NCBI37 [4]). CAST/EiJ and PWK/PhJ each have a high variation rate of approximately one variant per 130 bp with respect to the reference genome, and a slightly higher rate with respect to each other, while WSB/EiJ is more similar to the reference genome with approximately one variant per 375 bp. The genetic distance between these three strains make them ideal candidates for studying ASE, since we expect a large percentage of reads to contain distinguishing variants.

Table 1. Notation

$\mathbf{y}$	F1 k-mer profile. An $n \times 1$ vector where y_i indicates the number of times the i^{th} k-mer appears in the F1 sample
$\mathbf{z}^M, \mathbf{z}^P$	maternal and paternal k-mer profiles
$\mathbf{X}^M$	set of k-mer profiles of candidate transcripts from $\mathbf{z}^M$
$\mathbf{X}^P$	set of k-mer profiles of candidate transcripts from $\mathbf{z}^P$
$\mathbf{X}$	an $n \times m$ matrix equal to $[\mathbf{X}^M \cup \mathbf{X}^P]$, where n is number of k-mers and m is number of transcripts
$\mathbf{x}_j$	k-mer profile of the j^{th} candidate transcript
$x_{i,j}$	number of times the i^{th} k-mer occurs in the j^{th} candidate transcript
θ_j	estimated expression level for the j^{th} candidate transcript

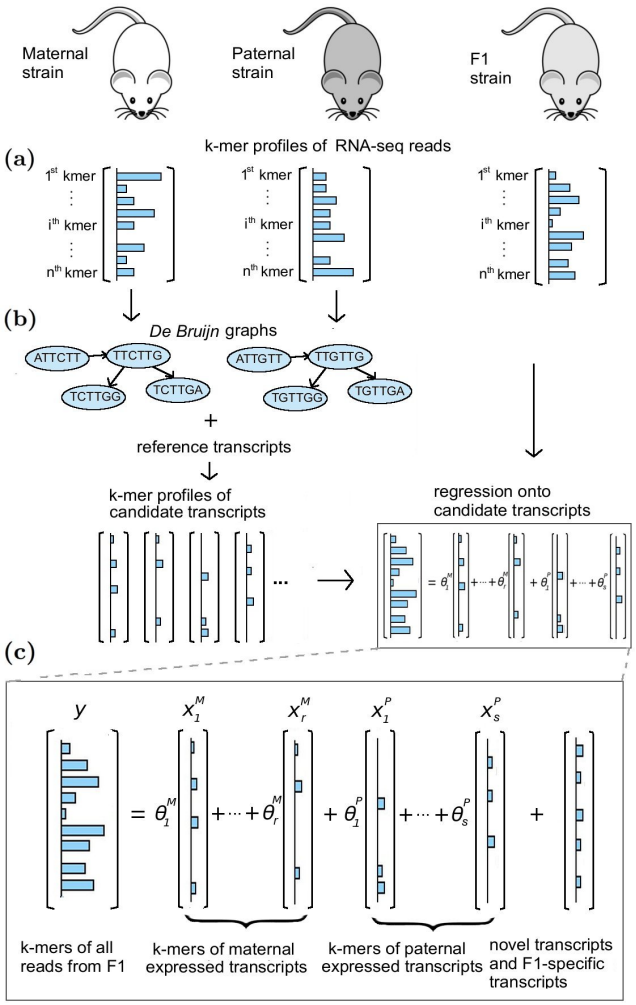

Fig. 1. Our pipeline for estimating allele-specific expression in F1 animals. **(a)** k-mer profiles are created for the maternal, paternal, and F1 strains, using all available RNA-seq reads from one sample of each strain. Each k-mer is also saved as its reverse complement, since we do not know the directionality of the read. **(b)** *De Bruijn* graphs are created for the maternal and paternal samples. Using annotated reference transcripts and the parental *De Bruijn* graphs, we select candidate transcripts which incorporate parental alleles from the *De Bruijn* graphs. **(c)** The k-mer profile of the F1 sample, y, is then regressed onto the candidate parental transcripts, $\{\mathbf{x}_1^M, \mathbf{x}_2^M, ...\mathbf{x}_r^M\} \cup \{\mathbf{x}_1^P, \mathbf{x}_2^P, ...\mathbf{x}_s^P\}$, and we estimate the expression level θ of each candidate transcript.

2 Approach

In this section, we discuss the parameters and assumptions of our proposed model and the underlying optimization problem.

2.1 Notation

Table (1) includes a description of the variables used in this paper. We denote the k-mer profiles of maternal candidate transcripts, $\mathbf{X}^M = \{\mathbf{x}_1^M, \mathbf{x}_2^M, ...\mathbf{x}_r^M\}$, and the k-mer profiles of paternal candidate transcripts, $\mathbf{X}^P = \{\mathbf{x}_1^P, \mathbf{x}_2^P, ...\mathbf{x}_s^P\}$, jointly as $\mathbf{X} = \mathbf{X}^M \cup \mathbf{X}^P$, a matrix representing the k-mer profiles of all candidate transcripts. Each candidate transcript k-mer profile is labeled as originating from the maternal k-mer profile, the paternal k-mer profile, or both if there are no differentiating variants between the parental k-mer profiles.

2.2 Regression model

We propose a modified lasso penalized regression model for estimating the abundance of each candidate transcript, with the assumption that the F1's k-mer profile y can be expressed as a linear combination of its expressed transcripts $\mathbf{X} = \{\mathbf{x}_1, ...\mathbf{x}_j, ...\mathbf{x}_m\}$ multiplied by their relative expression levels θ_j:

$$\mathbf{y} = \sum_{j=1}^{m} \theta_j \mathbf{x}_j. \tag{1}$$

To filter out non-expressed transcripts and prevent overfitting, each candidate transcript is penalized by an l_1-norm, parameterized by the regularization parameter λ and the inverse of w_j, where

$$w_j = median \begin{cases} \{z_i^M / x_{i,j}, \forall x_{i,j} > 0\}, & \mathbf{x}_j \in \mathbf{X}^M \\ \{z_i^P / x_{i,j}, \forall x_{i,j} > 0\}, & \mathbf{x}_j \in \mathbf{X}^P \\ \{(z_i^M + z_i^P) / x_{i,j}, \forall x_{i,j} > 0\}, & \mathbf{x}_j \in \mathbf{X}^P \cap \mathbf{X}^M \end{cases} \tag{2}$$

Therefore, transcripts that are expressed at a high level in the parental samples are more likely to be expressed at a high level in the F1 sample as well. Our objective function then becomes

$$\begin{aligned} \underset{\theta}{\text{argmin}} \quad & \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \theta_j x_{i,j})^2 + \lambda \sum_{j=1}^{m} \frac{\theta_j}{w_j} \\ \text{subject to} \quad & \theta_j \geq 0, \forall j, \end{aligned} \tag{3}$$

with each θ_j constrained to be nonnegative since they represent transcript expression levels.

3 Methods

3.1 Synthetic data

We used the Flux Simulator [9] to create simulated reads from the CAST/EiJ, PWK/PhJ, and CASTxPWK mouse genomes. We chose these two parental strains because they are well-annotated strains that differ significantly from the reference strain C57BL/6J and from each other. The transcript sequences for CAST/EiJ and PWK/PhJ were created using Cufflinks' gffread utility [27] with genomes from the Wellcome Trust Institute [13] and transcript annotation from the Ensembl Genome Database [3]. The positions from the reference transcript annotation files were updated with positions to the CAST/EiJ and PWK/PhJ genomes using MODtools [12].

We synthesized 10,000,000 100bp paired-end reads from both the CAST/EiJ and the PWK/PhJ genomes to represent reads from a maternal CAST/EiJ genome and a paternal PWK/EiJ genome. We specified the same set of 1000 transcripts with a positive number of expressed RNA molecules in both genomes. In addition, we merged two sets of 5,000,000 separately synthesized reads from both CAST/EiJ and PWK/PhJ to create a simulated F1 fastq file. From the merged CAST/EiJ and PWK/PhJ versions of transcript sequences, the Flux Simulator output 1156 unique transcripts sequences where at least 95% of the sequence is covered by reads, and we define this set of 1156 transcript sequences, representing 626 reference transcripts, as the truly expressed transcripts.

3.2 Real data

RNA from whole-brain tissues (excluding cerebellum) was extracted from 5 samples (CAST/EiJ female, PWK/PhJ male, WSB/EiJ male, CASTxPWK male and CASTxWSB female) using the Illumina TruSeq RNA Sample Preparation Kit v2. The barcoded cDNA from each sample was multiplexed across four lanes and sequenced on an Illumina HiSeq 2000 to generate 100 bp paired-end reads (2x100). This resulted in $2 \times 71,291,857$ reads for the CAST/EiJ sample, $2 \times 49,877,124$ reads for the PWK/PhJ sample, $2 \times 62,712,206$ reads for the WSB/EiJ sample, $2 \times 77,773,220$ reads for the CASTxPWK hybrid sample, and $2 \times 57,386,133$ reads for the CASTxWSB hybrid sample. Note that the selected samples were not true biological trios, but genetically equivalent. We also used the same female CAST/EiJ sample as the maternal model for both F1 hybrids.

3.3 Selecting candidate transcripts

We used a greedy approach for selecting candidate transcript sequences from the *De Bruijn* graphs of each parental k-mer profile. The k-mer size used for this and subsequent analyses was 32 bp. For each of the 93,006 reference transcripts provided by Ensembl [3], we match the reference transcript sequence to a path of k-mers in the *De Bruijn* graph, allowing for a maximum number of 5 mismatches within a sliding window of 25 bp, which is a sensible choice except in the case

of unusually dense SNPs or indels. In the case of mismatches, we replace the reference sequence with the sequence in the parental *De Bruijn* graph, thus creating updated candidate transcript sequences which reflect variants in the parental strains. If more than 80% of a transcript's k-mers are found in the *De Bruijn* graph, we consider it a candidate transcript. The k-mer profiles of the selected candidate transcript sequences are then used as features in our regularized regression model.

3.4 Coordinate descent

To optimize our objective function Eq. (3), we update θ_j using coordinate descent:

$$\begin{aligned} \theta_j =& \frac{max(\sum_{i=1}^{n} y_i^{(-j)} x_{i,j} - \frac{\lambda}{w_j}, 0)}{\|x_j\|_2^2}, \text{where} \\ y_i^{(-j)} =& y_i - \sum_{k \neq j} \theta_k x_{i,k}. \end{aligned} \tag{4}$$

Due to the high dimensional nature of our data (in real data, the number of k-mers, n, is approximately 5×10^7, and the number of candidate transcripts, m, is approximately 2×10^4), updating each θ_j on every iteration becomes inefficient. We therefore adapt the coordinate descent with a refined sweep algorithm as described by Li and Osher [18], where we greedily select to update only the θ_j that changes the most on every iteration. To save on computation per iteration, we can let $\beta_j = \sum_{i=1}^{n} y_i^{(-j)} x_{i,j}$ and precompute the matrix product $\mathbf{X}^T\mathbf{y}$, so that β can be updated at every iteration using only addition and a scalar-vector multiplication. The algorithm is described in Eq. (5), and proof of its convergence is provided by Li and Osher [18].

$$\begin{aligned} &\text{Initialize:} \\ &\quad \theta^0 = \mathbf{0} \\ &\quad \beta^0 = \mathbf{X}^T\mathbf{y} \\ &\quad \gamma = diag(\|\mathbf{x}_j\|_2^2) - \mathbf{X}^T\mathbf{X} \\ &\text{Iterate until convergence:} \\ &\quad \theta^* = \frac{max(\beta - \frac{\lambda}{\mathbf{w}}, 0)}{\|\mathbf{x}_j\|_2^2} \\ &\quad j = argmax|\theta^* - \theta^k| \\ &\text{Updates:} \\ &\quad \theta_j^{k+1} = \theta_j^* \\ &\quad \beta^{k+1} = \beta^k + \gamma_{j,:}(\theta_j^* - \theta_j^k) \\ &\quad \beta_j^{k+1} = \beta_j^k \end{aligned} \tag{5}$$

The coordinate descent algorithm terminates when the minimization objective Eq. (3) decreases by less than a threshold of 0.001 per iteration. For computational efficiency, the value of our objective function Eq. (3) is evaluated per τ iterations, where $\tau = 10^4$ initially. We decrease τ as the objective increases, until $\tau = 1$ for the final iterations. This saves significant computation time since the computation of the objective function contains a matrix multiplication and the regular updates do not, and the convergence of the algorithm is not affected as the updates are still being performed per iteration.

The lasso regularization parameter λ is chosen via 4-fold cross validation. It is important to note that the value of λ depends on the mean observed values for w_j, so different values of λ could be chosen for each trio.

4 Results

We analyzed a synthetic data set to ascertain the sensitivity and specificity of our estimation framework. We then applied our technique to two real data sets and evaluated them based on their ability to recapitulate known biological properties.

4.1 Synthetic data results

In our synthetic F1 sample, the Flux Simulator generated 1156 unique transcript sequences from both the maternal and paternal haplotypes with positive expression levels, representing 626 reference transcripts. We identified 4517 candidate parental transcript sequences from all reference mouse transcripts annotated by Ensembl, 1055 of which were truly expressed, representing 598 out of 626 truly expressed reference transcripts.

We selected the lasso regularization parameter λ to be 500 using 4-fold cross validataion. We took $\theta_j = 0$ to indicate transcript j was not expressed and calculated the sensitivity and specificity of our method in identifying which transcripts were expressed. For the chosen value of λ, we found the sensitivity to be 0.9553 (598/626) and the specificity to be 0.9880 (91278/92385).

Of the correctly identified expressed transcripts, the true and estimated expression levels had a Pearson correlation coefficient of 0.85, indicating high positive correlation, as shown in Fig. (S1). To allow for comparison of relative expression levels, we normalized both true and predicted expression levels to have a mean value of 1 across all expressed transcripts. The mean absolute error between true and predicted expression levels was 0.3128 for the chosen value of λ. True positive rates, false positive rates, and mean absolute error of predicted expression levels for different values of λ are summarized in Fig. (2).

Among the 598 correctly identified expressed transcripts, 544 had differentiable paternal and maternal candidate sequences. Of these, 141 exhibited ASE, as defined by having a maternal contribution ratio (maternal expression level divided by total expression level) outside the range $[0.4, 0.6]$. Our model correctly

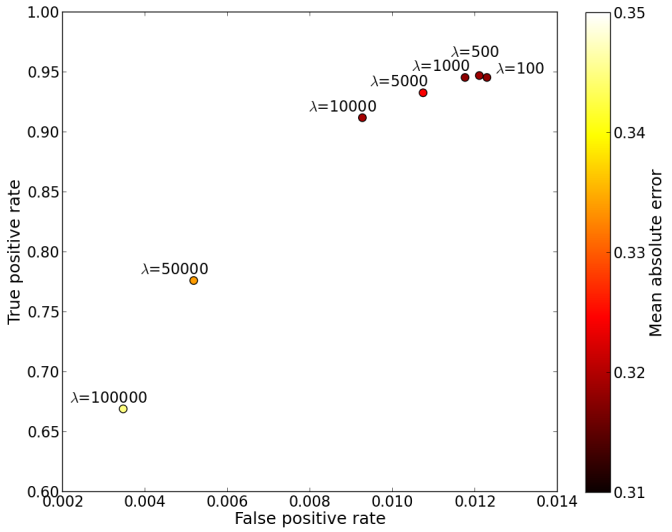

Fig. 2. True positive rate vs. false positive rate for different values of λ. Each point is colored by the mean absolute error between normalized true and estimated expression levels for all transcripts correctly classified as expressed.

identified 109 transcripts exhibiting ASE and correctly rejected 293 transcripts not exhibiting ASE, achieving a sensitivity of 0.77 and specificity of 0.73.

We compared our results with Trinity [7], since its *de novo* assembly methods are able to separate maternal and paternal versions of transcripts better than reference alignment-based methods.

To assemble candidate transcripts from the maternal and paternal strains, we ran Trinity with its default parameters on the synthetic maternal CAST/EiJ and paternal PWK/PhJ samples. Per Trinity's downstream analysis guidelines, we then aligned reads from the synthetic F1 sample to the assembled parental transcript sequences using Bowtie [14] then estimated expression levels using RSEM [16].

Trinity assembled 4215 transcript sequences from both parental strains. Following their guidelines to eliminate false positives, we retained 3336 transcript sequences representing at least 1% of the per-component expression level. We used a criterion of Levenshtein distance less than 10% of the true transcript length to match annotated transcripts to the *de novo* transcripts sequences reported by Trinity. With this criterion, only 110 out of 626 truly expressed transcripts were present in the set of expressed transcripts found by Trinity. In this set, the mean Levenshtein distance from each true transcript sequence to the Trinity sequences was 0.12% of the true transcript length, with the maximum distance being 2.6% of the true transcript length, suggesting our matching criterion of 10% Levenshtein distance was generous.

Out of the 110 assembled transcripts correctly identified, 81 had nonzero expression levels, making the sensitivity for baseline expression detection 0.13. However, of the 81 correctly identified transcripts, the Trinity-Bowtie-RSEM pipeline produced a high correlation of 0.88 between true and estimated expression levels.

Of the 81 expressed transcripts correctly identified by Trinity, 63 originated from reference transcripts with ASE. Trinity correctly identified 20 true positives and 16 true negatives, with a sensitivity of 0.32 and specificity of 0.89.

4.2 Real data results

Table 2. Dimensions and Results from Real Data

	CASTxPWK	**CASTxWSB**
k-mers in merged trio k-mer profile	118,100,824	118,383,117
k-mers in candidate transcripts	42,688,910	52,715,089
k-mers in estimated expressed transcripts	42,482,315	52,162,586
candidate transcripts	23,585	29,155
estimated expressed transcripts	17,118	20,596
candidate genes	7,393	8,532
estimated expressed genes	7,148	8,242
expressed genes with isoforms from both parents	4,065	5,183

We applied our methods to a male CASTxPWK F1 sample and a female CASTxWSB F1 sample. We first created *De Bruijn* graphs for a CAST/EiJ female, a PWK/EiJ male, and a WSB/EiJ male, representing the parental *De Bruijn* graphs of our two F1 samples. To eliminate erroneous reads in each strain, we filtered k-mers appearing fewer than 5 times. Using Algorithm 2, we selected 15,287 candidate transcripts from the CAST/EiJ *De Bruijn* graph, 9,852 candidate transcripts from the PWK/EiJ graph, and 16,023 candidate transcripts from the WSB/EiJ graph. For each F1 sample, transcript sequences without differentiating variants between the two parental strains were merged into a single candidate transcript. This resulted in 23,585 candidate transcripts for CASTxPWK and 29,155 candidate transcripts for CASTxWSB, representing 7,393 and 8,532 candidate genes, respectively.

The CAST/EiJ, PWK/EiJ and CASTxPWK trio had a merged k-mer profile of 118,100,824 k-mers, 42,688,910 (36.1%) of which appeared in our candidate transcripts. Similarly, the CAST/EiJ, WSB/EiJ and CASTxWSB trio had a merged k-mer profile of 118,383,117 k-mers, 52,715,089 (44.5%) of which appeared in its set of candidate transcripts. We verified most the k-mers in the F1 samples not appearing in candidate transcripts have few occurrences. The k-mers with high profiles which do not appear in candidate transcripts are most likely due to poly(A) tails, transcripts with dense variants in the parental strains, or transcripts expressed by the F1 strain but not the parents, as shown in Fig. (S2)

Using the penalty parameter $\lambda = 10^4$ for both F1 samples, our methods found 17,118 non-zero θ values in the CASTxPWK sample and 20,596 non-zero θ values in the CASTxWSB sample, corresponding to as many estimated expressed transcripts. This represented 7,148 of 7,393 and 8,242 of 8,532 estimated expressed genes, respectively. These results are summarized in Table (2). We estimated the expression level of each gene by summing the θ values for all expressed isoforms, both maternal and paternal, of each gene.

To assess our ability to estimate ASE, we looked at the maternal contribution ratio of all expressed genes with candidate isoforms from both parents and differentiating variants between the two parents. Maternal contribution ratio of a gene is defined as the ratio of the expression levels from all maternal isoforms to the expression levels from both paternal and maternal isoforms of the gene. The distribution of maternal contribution ratios for both F1 samples is shown in Fig. (3). The median maternal contribution ratio for both the male CASTxPWK sample and the female CASTxWSB sample is around 0.5, as expected. In the male CASTxPWK sample, a higher number of genes are maternally expressed, which is expected since genes on the X chromosome and mitochondria should be maternally expressed in males. We verified several genes that are known to exhibit ASE [8] [29] as having high maternal contribution ratios if maternally expressed and low maternal contribution ratios if paternally expressed.

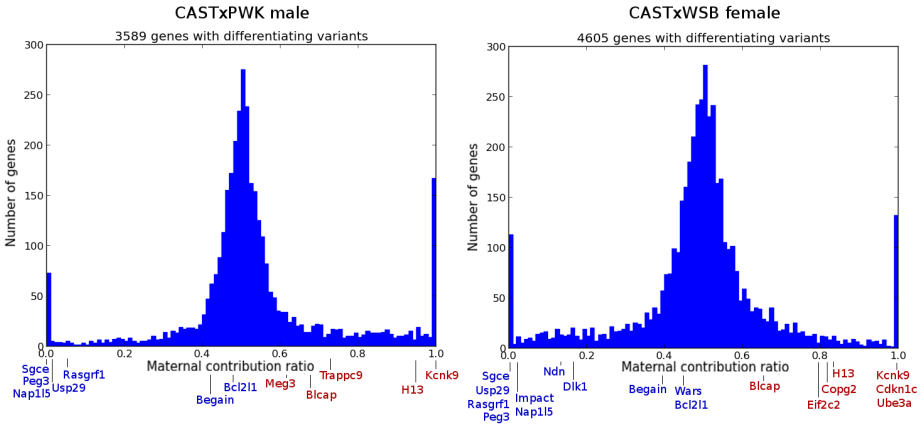

Fig. 3. Histogram of the maternal contribution ratios of all expressed genes with candidate isoforms from both parental strains and containing differentiating variants between the parental strains. On the bottom of each plot, several genes known to be maternally expressed in literature are highlighted in red, and several genes known to be paternally expressed are highlighted in blue.

In addition, we examined the maternal contribution ratios of all expressed genes on the X chromosome with candidate isoforms from both parents and differentiating variants between the parents. In the male CASTxPWK sample,

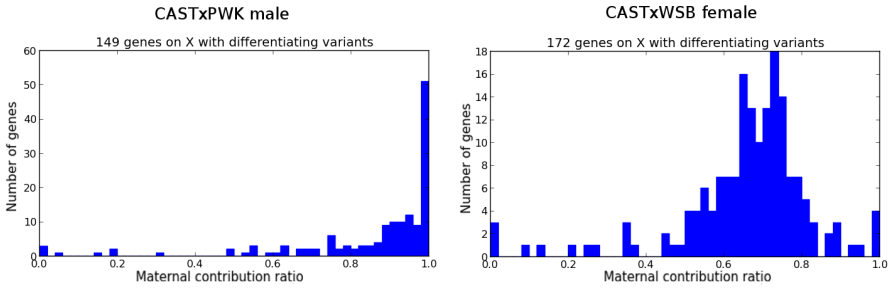

Fig. 4. Histogram of the maternal contribution ratio of all expressed genes on the X chromosome with candidate isoforms from both parental strains and containing differentiating variants between the parental strains. In the male CASTxPWK sample, the median maternal contribution ratio is 0.94. In the female CASTxWSB sample, the median maternal contribution ratio is 0.68. Both are in the expected range of maternal contribution ratio of X-chromosome genes in male and female animals, respectively.

we expect all genes on the X chromosome to be maternally expressed, since its X chromosome is inherited from the maternal strain. In the female CASTxWSB sample, we expect most genes on X to be expressed with a 0.6-0.7 maternal contribution ratio due to a known maternal bias in X inactivation [28] [2]. As expected, we found the median maternal contribution ratio to be 0.94 in the male CASTxPWK sample and 0.68 in the female CASTxWSB sample. The distributions of maternal contribution ratios of genes on the X chromosome are plotted in Fig. (4).

4.3 Speed and Memory

We ran our methods on a single 1600 MHz processor on a machine with 32 GB RAM. The *De Bruijn* graphs of our samples take up around 1GB of disk space. The selection of candidate transcripts takes approximately 2-3 hours per parental strain, and the coordinate descent algorithm converges after approximately 1 to 3 million iterations, which takes around 1-2 hours on our machine. We were able to take advantage of the sparseness of our candidate transcript k-mer profile matrix $\mathbf{X}$ by storing them as sparse matrices using the Scipy.sparse package.

5 Discussion

We have developed methods to estimate expression levels for maternal and paternal versions of transcripts from RNA-seq trio data. Our need for such methods arose when we realized that although we have RNA-seq data of many biological trios and wish to analyze ASE of F1 strains, current methods, both alignment-based and *de novo*, do not include standard pipelines that take advantage of available RNA-seq data from parental strains. Our model is able to exploit the information from the maternal and paternal RNA-seq reads and build candidate

transcripts that accurately reflect the F1 strain's transcriptome, and it does so without requiring a database of variants of the parental strains. Our proposed methods still rely on the existence of an annotated reference transcriptome, which is essential for making biologically meaningful observations.

Our methods performed well when compared to a Trinity-Bowtie-RSEM pipeline, which incorporates a state-of-the-art *de novo* assembler and aligner. We were able to achieve high sensitivity and specificity (0.9553 and 0.9883) in detecting baseline expression of transcripts. Of the correctly identified expressed transcripts, we were also able to correctly identify more transcripts exhibiting ASE, with a sensitivity of 0.77, compared to Trinity's low ASE sensitivity of 0.32. The pipeline we used with Trinity also made use of parental RNA-seq data, since we separately assembled transcript sequences from maternal and paternal reads, then aligned the F1 reads to the entire set of assembled transcript sequences. However, Trinity still had a low sensitivity of 0.13 for determining baseline expression, since the main challenge we faced using Trinity was mapping the assembled sequences back to known reference transcripts.

The dimensionality of our data can be large. In our real data, we have approximately 5×10^7 k-mers after filtering and tens of thousands of candidate transcripts. Despite the high dimensionality of our k-mer space and transcripts space, we were able to use a refined coordinate descent algorithm to efficiently perform lasso regression. Although not implemented, we could also decrease our k-mer space without affecting the solution by merging overlapping k-mers into contigs.

Since our candidate transcripts are generated from annotated reference transcripts, our methods do not currently assemble novel transcript sequences. However, it is possible to model the k-mer profiles of all novel transcripts as the residual of our linear regression, and *de novo* assembly of the residual k-mers could then generate sequences of novel transcripts. Another limitation of our model lies in its inability to detect genes exhibiting overdominance, where the expression level is high in the F1 animal but nonexistent in the parental strains. This could be remedied by also selecting candidate transcripts from the F1 *De Bruijn* graph itself to add to our feature space.

The strength of our methods lies in the ability to determine ASE directly from RNA-seq data in diploid trios without prior knowledge of genomic variation in the parental genomes. This straightforward regression approach is tolerant of imbalanced read counts in different samples, as demonstrated by our reasonable maternal contribution ratio distribution in the male CASTxPWK F1 sample (Fig 3), despite the CAST/EiJ read count being nearly 1.5 times as high as the PWK/EiJ read count. Our methods could even be extended to ascertain ASE in any animal that is a hybrid of two or more isogenic ancestral genomes, such as the recombinant inbred strains often used as genetic reference panels.

6 Acknowledgements

The authors thank Fernando Pardo-Manuel de Villena and the UNC Center for Excellence in Genome Sciences (NIH P50 MH090338) for providing the sequencing data used in this study.

References

1. Michael Burrows and David J Wheeler. *A block-sorting lossless data compression algorithm*. Citeseer, 1994.
2. Lisa Helbling Chadwick, Lisa M Pertz, Karl W Broman, Marisa S Bartolomei, and Huntington F Willard. Genetic control of x chromosome inactivation in mice: definition of the xce candidate interval. *Genetics*, 173(4):2103–2110, 2006.
3. A.T. Chinwalla, L.L. Cook, K.D. Delehaunty, G.A. Fewell, L.A. Fulton, R.S. Fulton, T.A. Graves, L.D.W. Hillier, E.R. Mardis, J.D. McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.
4. Deanna M Church, Leo Goodstadt, LaDeana W Hillier, Michael C Zody, Steve Goldstein, Xinwe She, Carol J Bult, Richa Agarwala, Joshua L Cherry, Michael DiCuccio, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, 7(5):e1000112, 2009.
5. N.G. de Bruijn and P. Erdos. A combinatorial problem. *Koninklijke Netherlands: Academe Van Wetenschappen*, 49:758–764, 1946.
6. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
7. M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.
8. Christopher Gregg, Jiangwen Zhang, Brandon Weissbourd, Shujun Luo, Gary P Schroth, David Haig, and Catherine Dulac. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *science*, 329(5992):643–648, 2010.
9. Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083, 2012.
10. M. Guttman, M. Garber, J.Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M.J. Koziol, A. Gnirke, C. Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5):503–510, 2010.
11. Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.
12. Shunping Huang, Chia-Yu Kao, Leonard McMillan, and Wei Wang. Transforming genomes using mod files with applications. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM*, 2013.
13. T.M. Keane, L. Goodstadt, P. Danecek, M.A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, 2011.

14. B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.
15. V.I. Levenshtein. Binary codes capable of correcting deletions, insertions. Technical report, and reversals. Technical Report 8, 1966.
16. B. Li and C.N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.
17. W. Li, J. Feng, and T. Jiang. Isolasso: a lasso regression approach to rna-seq based transcriptome assembly. *Journal of Computational Biology*, 18(11):1693–1707, 2011.
18. Y. Li and S. Osher. Coordinate descent optimization for l-1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Probl. Imaging*, 3(3):487–503, 2009.
19. Ruijie Liu, Ana-Teresa Maia, Roslin Russell, Carlos Caldas, Bruce A Ponder, and Matthew E Ritchie. Allele-specific expression analysis methods for high-density snp microarray data. *Bioinformatics*, 28(8):1102–1108, 2012.
20. Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
21. G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S.D. Jackman, K. Mungall, S. Lee, H.M. Okada, J.Q. Qian, et al. De novo assembly and analysis of rna-seq data. *Nature methods*, 7(11):909–912, 2010.
22. James Ronald, Joshua M Akey, Jacqueline Whittle, Erin N Smith, Gael Yvert, and Leonid Kruglyak. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Research*, 15(2):284–291, 2005.
23. J. Rozowsky, A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, et al. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7(1), 2011.
24. D.A. Skelly, M. Johansson, J. Madeoy, J. Wakefield, and J.M. Akey. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from rna-seq data. *Genome research*, 21(10):1728–1737, 2011.
25. C. Trapnell, L. Pachter, and S.L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.
26. C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *nature protocols*, 7(3):562–578, 2012.
27. C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
28. Xu Wang, Paul D Soloway, Andrew G Clark, et al. Paternally biased x inactivation in mouse neonatal brain. *Genome Biol*, 11(7):R79, 2010.
29. Xu Wang, Qi Sun, Sean D McGrath, Elaine R Mardis, Paul D Soloway, and Andrew G Clark. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PloS one*, 3(12):e3839, 2008.
30. H. Yang, J.R. Wang, J.P. Didion, R.J. Buus, T.A. Bell, C.E. Welsh, F. Bonhomme, A.H.T. Yu, M.W. Nachman, J. Pialek, et al. Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics*, 43(7):648–655, 2011.

A Supplemental Figures

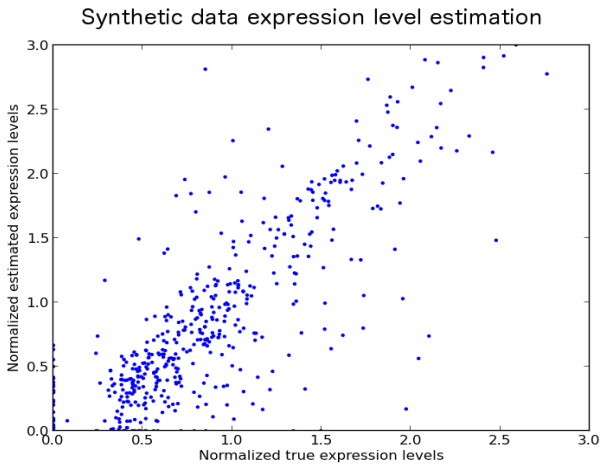

Fig. S1. Predicted versus actual expression levels from synthetic data. Expression levels were normalized to have a mean value of 1.The Pearson correlation coefficient is 0.85 among the 1055 correctly identified expressed transcript sequences.

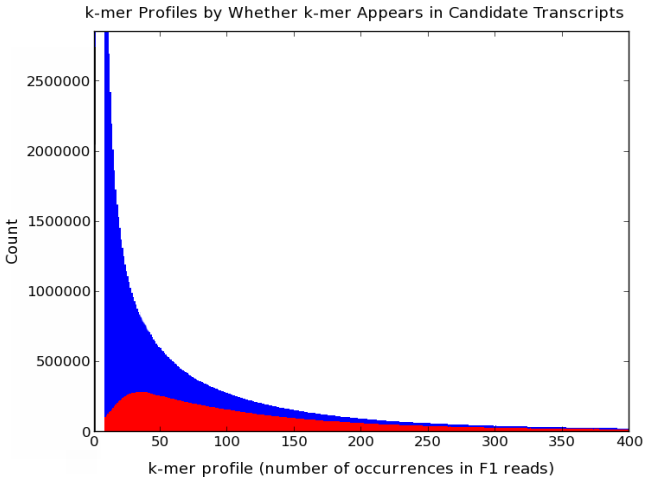

Fig. S2. Stacked histogram of k-mers in the real CASTxPWK k-mer profile, sorted by the number of times each k-mer appears in the F1 reads. K-mers appearing in candidate transcripts are in red, and k-mers not appearing in candidate transcripts are in blue. The majority of k-mers not appearing in candidate transcripts have low number of occurrences, suggesting they are from lowly expressing genes or erroneous reads.